УДК 004.048

И.С. Сальников, С.В. Терещенко
Институт проблем искусственного интеллекта МОН Украины и НАН Украины, г. Донецк
Украина, 83050, г. Донецк, ул. Артёма, 118б

Алгоритм внесения роботизированным компьютером информации из бумажных документов в электронные базы данных

I.S. Salnikov, S.V. Tereshchenko
Institute of Artificial Intelligence, Donetsk, Ukraine
Ukraine, 83050, Donetsk, Artema st., 118b

Algorithm of Entering Information from Paper Documents into Electronic Databases by Robotic Computer

І.С. Сальников, С.В. Терещенко
Інститут проблем штучного інтелекту МОН України і НАН України, м. Донецьк
Україна, 83050, м. Донецьк, вул. Артема, 118б

Алгоритм внесення роботизованим комп’ютером інформації з паперових документів в електронні бази даних

В этой работе рассматривается процесс обработки бумажной документации. Предложены алгоритм автоматизации этого процесса с помощью роботизированного компьютера и возможная структура электронных документов.
Ключевые слова: роботизированный компьютер, бюстер, алгоритм, автоматизация.

In this work considers the processing of paper documents. Are proposed algorithm to automate this process with use of a robotic computer and the possible structure of electronic documents.
Key words: robotic computer, buster, algorithm, automation.

У цій праці розглядається процес обробки паперової документації. Запропоновані алгоритм автоматизації цього процесу за допомогою роботизованого комп'ютера та можлива структура електронних документів.
Ключові слова: роботизований комп'ютер, бюстер, алгоритм, автоматизація.

Перенос информации с бумажных источников в электронные базы данных предприятий является одним из наиболее важных процессов в документообороте. Актуальность автоматизации этого процесса обусловлена несколькими факторами, наиболее важным из которых является обеспечение достоверности введенных данных. Вторым фактором является необходимость обеспечения высокой скорости обработки документов, от которой зависит эффективность работы предприятия. Производительность оператора снижается с каждым обработанным им документом. Глаза устают, когда приходится долго смотреть на экран компьютера и это сказывается на всем организме. В частности, замедляются рефлексы и ослабляется внимание [1]. Как следствие – оператор начинать работать медленнее и допускать больше ошибок. Третьим фактором является неравномерность объемов поступающих на обработку документов в разные периоды времени. Чтобы справляться с пиковыми нагрузками предприятиям приходиться содержать большой штат сотрудников, который большую часть времени работает в неполную силу.

При обработке документов человек ловит взглядом характерное словосочетание и сразу понимает, к чему оно относится и как следует обрабатывать эту информацию. В то же время для робота это весьма непростая задача. Одной из важнейших задач на пути к разработке роботизированного компьютера, способного заменить на рабочем месте офисного сотрудника, является распознавание образов. Роботизированный компьютер должен не только распознавать множество объектов окружающего мира, но и уметь их классифицировать.

Целью данной работы является разработка алгоритма внесения роботизированным компьютером информации из бумажных документов в электронные базы данных и формирование структуры электронного документа.

Алгоритм внесения роботизированным компьютером информации из бумажных документов в электронные базы данных

Любой тип документа имеет перечень обязательных полей, которые необходимы для проведения каких-либо операций над документом. Также в документе могут содержаться дополнительные поля, определенные спецификой работы конкретного предприятия. Наименования полей выступают в качестве ключевых слов (дескрипторов). Ключевые слова, которые являются эквивалентными либо синонимами, объединяются в одну дескрипторную группу.

При анализе документа человек опирается на характерные для данного типа документов ключевые слова и словосочетания. Например, на рис. 1 встречаются ключевые слова: «Постачальник», «Одержувач», «Сплатити до» и пр.

Образец документа

Постачальник: ООО "Би Энд Пи, ЛТД (совместная деятельность)"
Р/р 260083012307 в ГОУ ПИБ в г.Киеве, МФО 300012
ІПН , номер свідоцтва 36073005
код ЕДРПОУ: 14351789, Адреса: пр. Червонозоряний, 3,

Одержувач: ДП "Квіза-Трейд"
Сплатити до: 02.09.2004, тел.:

Рахунок-фактура №6

від 30 серпня 2004 р.

№	Назва	Од. вим.	Кіл.	Ціна без ПДВ	Сума без ПДВ
1	Вода мин. Неаполис 0,33л	шт	3	22,00	66,00
2	Драже Свиточ изюм в какао 80г	шт	2	12,00	24,00
3	Арахис Клинское пиво копчений 30г	шт	1	22,00	22,00
				Разом без ПДВ:	**112,00**
				Знижка/Надбавка:	**0,00**
				Разом без ПДВ:	**112,00**
				ПДВ:	**22,40**
				Всього з ПДВ:	**134,40**

Всього на суму:
Сто тридцять чотири грн 40 коп.
ПДВ: 22,4 грн.

Рисунок 1 – Отсканированный образец документа на экране бюстера

Тип документа человек определяет на базе определенного дескриптора («Рахунок-фактура» на рис. 1) либо исходя из структуры документа (перечня ключевых слов, характерного для определенного типа документов). Значение дескриптора (поля документа) обычно располагается либо справа от него, либо снизу.

Анализ документа роботизированный компьютер может выполнять на базе определенных дескрипторов, которые он может получить как из сети от другого компьютера, так и непосредственно от руководителя.

На рис. 2 представлен алгоритм, который позволит роботизированному компьютеру (бюстеру) обрабатывать документы аналогично тому, как это делал бы человек.

Документы, предназначенные на обработку, помещаются в очередь. В данном случае под очередью подразумевается некоторая ячейка (контейнер), в которую будут складывать документы на бумажных носителях. Роботизированный компьютер через блок управления механическими устройствами будет подавать команды манипулятору для перемещения документов из очереди на обработку, листания документов, переворота листов документа и перемещения обработанных документов в соответствующую ячейку [3].

Ячейку необработанных документов целесообразно снабдить детектором веса. Таким образом, как только в ячейку будет помещен какой-либо документ – детектор веса отправит сигнал о том, что в очереди появился документ для обработки. Закончив обработку документа, роботизированный компьютер отправляет детектору веса запрос на наличие в очереди документов для обработки. Если детектор веса возвращает нулевое значение, то компьютер через определенные интервалы времени отсылает ему запросы на наличие документов в очереди.

Документы могут состоять из нескольких скрепленных листов. Поэтому должен быть предусмотрен механизм листания документа.

Документ, изъятый из ячейки необработанных документов, сканируется с обеих сторон. Если документ состоит из нескольких скрепленных листов, то все они сканируются с обеих сторон.

Для распознавания отсканированного документа возможно использование существующих систем интеллектуального (Intelligent Character Recognition, ICR) распознавания символов, которые часто используются для распознавания как печатных, так и рукописных текстов. Примерами систем, причисляющих себя к категории ICR, являются: FineReader, OmniPage Professional, Readiris Corporate, Type Reader Desktop [4].

В результате распознавания текста формируется многомерный массив лексем. То есть каждая строка документа будет представлять собой массив текстовых значений. Каждая лексема представляет собой пару значений: непосредственно сама лексема и координата по оси X. Хранение информации о координатах по оси X необходимо при поиске значения дескриптора.

Согласно алгоритму, представленному на рис. 2, в массиве лексем производится поиск дескриптора, определяющего тип документа. Если он не найден, то из массива лексем выбираются все лексемы, которые совпадают с дескрипторами типовых документов. Затем подсчитывается количество совпадений элементов массива дескрипторов документа с элементами массивов дескрипторов типовых документов. Чем выше процент совпадений с дескрипторами определенного типа документов, тем больше вероятность, что к этому типу относится отсканированный документ. Если процент совпадений одинаков либо незначительно отличается у нескольких типов документов, то производится сравнение последовательности появления дескрипторов в отсканированном документе и в типовых документах (построчно сверху вниз). Соответственно, чем больше совпадений в порядке следования дескрипторов отсканированного документа с типовым, тем больше вероятность, что он относится к этому типу.

Если типовой массив дескрипторов отсканированного документа не совпал ни с одним массивом типового документа, то необходимо проверить ячейку необработанных документов (очередь) на наличие продолжения обрабатываемого документа (поскольку документ может состоять из нескольких отдельных листов). Если в очереди нет продолжения этого документа, следует отложить документ и отправить запрос на получение дополнительной информации в сеть и руководителю.

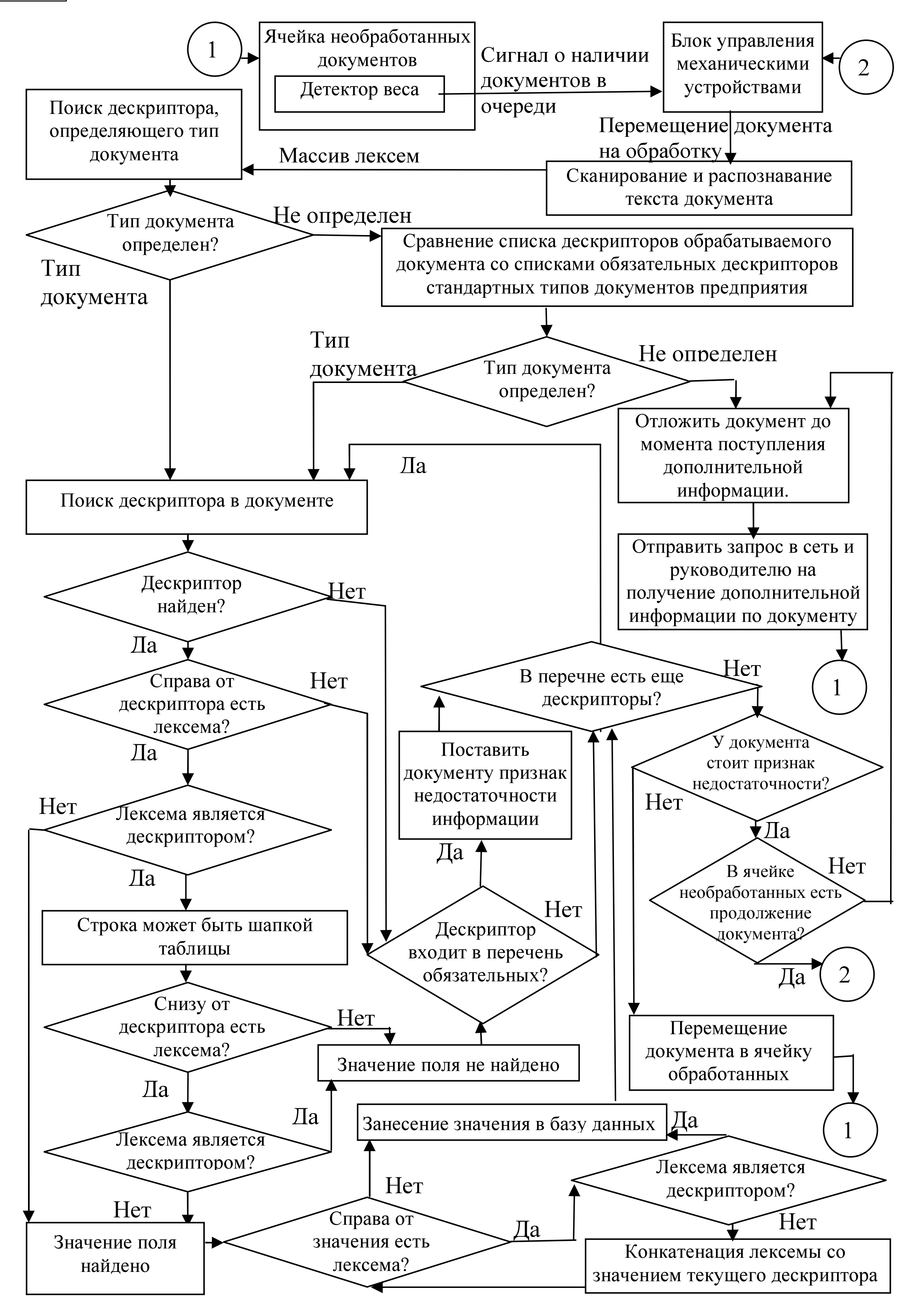


Рисунок 2 – Алгоритм обработки бумажных документов бюстером

Если тип документа определен – начинается его обработка, то есть занесение содержащейся в нем информации в базу данных. Каждый дескриптор конкретного типа документа ищется в массиве лексем отсканированного документа. Значение дескриптора ищется вначале справа (то есть в той же строке, что и дескриптор, но правее). Если лексема, стоящая справа от дескриптора не является также дескриптором, то это и есть искомое значение, которое заносится в базу данных. Иначе поиск значения осуществляется на следующей строке непосредственно под дескриптором. То, что лексема может являться значением дескриптора, расположенного над ней определяется на основе сравнения ее координат по оси X с координатами дескриптора с учетом некоторого среднеквадратического отклонения. Если лексема под дескриптором не является также дескриптором, то значение найдено.

При занесении значения дескриптора в базу знаний производится проверка соседней справа лексемы. Если она не является дескриптором, то она конкатенируется со значением обрабатываемого дескриптора. Конкатенация лексем происходит до тех пор, пока справа от лексем не встречается новый дескриптор. Если последняя в строке лексема не является дескриптором, то возможна конкатенация с лексемами следующей строки.

Если и справа и снизу от дескриптора стоят только дескрипторы – это значит, что в документе не заполнено значение для этого поля. В таком случае необходимо проверить, относится ли этот дескриптор к перечню обязательных для данного типа документов. Если нет, то можно просто перейти к обработке следующего дескриптора. Если же этот дескриптор относится к перечню обязательных, то документу ставится признак недостаточности информации, после чего производится поиск значения следующего дескриптора в массиве лексем.

Если в обрабатываемом документе были найдены не все обязательные дескрипторы, то необходимо проверить наличие его продолжения в ячейке с необработанными документами. Для этого вначале отсылается запрос детектору веса. Если он возвращает нулевое значение, то ячейка пуста и продолжения документа нет. Если значение ненулевое, то блок управления механическими устройствами подает команду манипулятору извлечь следующий документ из ячейки. Документ сканируется, распознается и проверяется, не является ли он продолжением предыдущего документа.

Если по окончании обработки всех дескрипторов у документа стоит признак недостаточности информации, то он помещается в электронную очередь документов и отправляется запрос на получение по нему дополнительной информации в сеть и руководителю. Несмотря на отсутствие какой-то части информации, данные из этого документа заносятся в базу данных. Если у документа заполнены все обязательные поля, то его обработка завершена.

Структура электронного документа

На рис. 3 представлена общая структура электронного документа, которая состоит из 5 связанных таблиц: «тип документа», «документ», «тело документа», «дескриптор», «тип документа / дескриптор».

Одинаковые дескрипторы могут находиться в различных типах документов. При этом они могут быть обязательными для одного типа документа и необязательными для другого. Также они могут иметь различный порядок расположения (вес дескриптора) в теле документа. Если вес дескриптора равен нулю, то порядок его расположения в теле документа определенного типа не имеет значения. Если вес

дескриптора отличен от нуля, то порядок его расположения должен быть в соответствии с его весом. Чем меньше вес дескриптора, тем выше он должен быть расположен в теле документа относительно дескрипторов с большим весом. Дескрипторы с одинаковым весом могут располагаться в любой последовательности относительно друг друга.

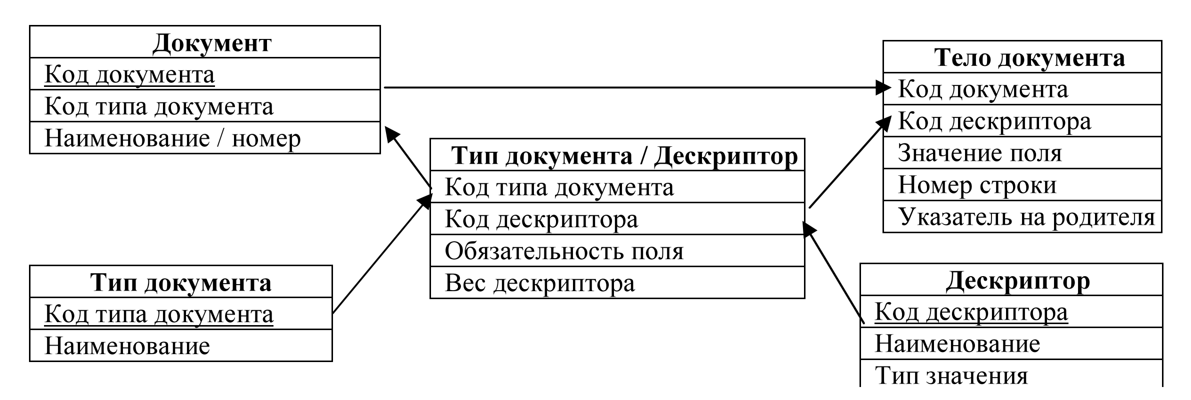


Рисунок 3 – Общая структура электронного документа

Номер строки – это номер строки отсканированного документа, в которой находится значение дескриптора. Указатель на родителя необходим для структурирования информации, хранимой в документе. Если значение дескриптора находится в таблице, то указателем на родителя будет номер строки, в которой находится шапка соответствующей таблицы. Если значение дескриптора не находится в таблице, то указатель на родителя будет равен нулю (то есть будет ссылаться на сам документ).

Рассмотрим выполнение бюстером типичной складской операции – внесение в базу данных информации о поступившем документе (рис. 1). В результате распознавания его текста формируется многомерный массив лексем, представленный в табл. 1. В таблице лексемы разделены между собой пробелами и знаками препинания.

Таблица 1 – Массив лексем обрабатываемого документа

Номер строки	Массив лексем строки обрабатываемого документа
1	Постачальник ООО 'БиЭндПи, ЛТД (совместная деятельность)'
2	Р/р 260083012307 в ГОУ ПИб в г.Киеве, МФО 300012
3	ІПН, номер свідоцтва 36073005
4	код ЄДРПОУ 14351789, Адреса пр Червонозоряний, 3
5	Одержувач ДП 'Квіза-Трейд'
6	Сплатити до 02.09.2004 тел
7	Рахунок-фактура №6
8	від 30 серпня 2004р.
9	№ Назва Од. вим. Кіл. Ціна без ПДВ Сума без ПДВ
10	1 Вода мин. Неаполис 0,33л шт 3 22.00 66.00
11	2 Драже Свиточ изюм в какао 80г шт 2 12.00 24.00
12	3 Арахис Клинское пиво копченый 30г шт 1 22.00 22.00
13	Разом без ПДВ: 112,00
14	Знижка/Надбавка: 0,00
15	Разом без ПДВ: 112.00
16	ПДВ: 22,80
17	Всього з ПДВ: 134,40
18	Всього на суму:
20	Сто тридцять чотири грн 40 коп.
21	ПДВ: 22.4 грн

В представленном массиве присутствует лексема «Рахунок-фактура», которая определяет тип документа. Предположим, что у некого предприятия для этого типа документов определен перечень дескрипторов, представленный в табл. 2. Дескрипторы, входящие в одну дескрипторную группу представлены через точку с запятой.

Таблица 2 – Дескрипторные группы типа документа «Рахунок-фактура»

Код дескриптора	Дескриптор	Обязательность дескриптора	Вес дескриптора
Д1	Постачальник	Да	1
Д2	Одержувач	Да	1
Д3	Сплатити; сплатити до	Да	1
Д4	Назва	Да	2
Д5	Од. вим.; одиниця виміру	Да	2
Д6	Кіл.; кількість	Да	2
Д7	Ціна без ПДВ	Да	3
Д8	Сума без ПДВ	Да	4
Д9	ПДВ	Да	4
Д10	Всього з ПДВ	Да	5
Д11	тел.; телефон	Нет	0

Таблица 3 – Тело документа

Код документа	Код дескриптора	Значение поля	Номер строки	Указатель на родителя
РФ6	Д1	ООО 'БиЭндПи, ЛТД (совместная деятельность)' Р/р 260083012307 в ГОУ ПИб в г.Киеве, МФО 300012 ІПН, номер свідоцтва 36073005 код ЄДРПОУ 14351789, Адреса пр Червонозоряний, 3	1	0
РФ6	Д2	ДП 'Квіза-Трейд'	5	0
РФ6	Д3	02.09.2004	6	0
РФ6	Д4	Вода мин. Неаполис 0,33л	10	9
РФ6	Д5	шт	10	9
РФ6	Д6	3	10	9
РФ6	Д7	22.00	10	9
РФ6	Д8	66.00	10	9
РФ6	Д4	Драже Свиточ изюм в какао 80г	11	9
РФ6	Д5	шт	11	9
РФ6	Д6	2	11	9
РФ6	Д7	12.00	11	9
РФ6	Д8	24.00	11	9
РФ6	Д4	Арахис Клинское пиво копченый 30г	12	9
РФ6	Д5	шт	12	9
РФ6	Д6	1	12	9
РФ6	Д7	22.00	12	9
РФ6	Д8	22.00	12	9
РФ6	Д9	22,80	16	9
РФ6	Д10	134,40	17	9

В массиве лексем обрабатываемого документа ведется поиск каждого дескриптора из перечня дескрипторов соответствующего типа документа. В первой строке массива лексем находится ключевое слово «Постачальник». Согласно алгоритму обработки документов (рис. 2) проверяем соседнюю справа лексему («ООО»). Она не входит в перечень дескрипторов, а значит, является либо значением, либо частью значения дескриптора «Постачальник». Далее по алгоритму проверяем лексему

справа («БиЭндПи»). Аналогичным образом до конца строки проверяем лексемы справа. Ни одна из них не входит в перечень дескрипторов, то есть их значения следует конкатенировать. Лексема в начале следующей строки входит в перечень дескрипторов («Р/р»), поэтому конкатенированные значения предыдущих лексем сохраняются как значение дескриптора «Постачальник».

Аналогичным образом заполняем значения дескрипторов «Р/р», «МФО», «ІПН, номер свідоцтва», «ЄДРПОУ», «Адреса», «Одержувач », «Сплатити до». Ни справа, ни снизу от дескриптора «тел.» нет лексемы, которая бы не входила в перечень дескрипторов. Но этот дескриптор не является обязательным, поэтому его заполнение пропускаем и переходим к заполнению значения следующего дескриптора – «Назва». Справа от него стоит лексема, которая входит в перечень дескрипторов. Проверяем лексемы до конца строки. Все они входят в перечень дескрипторов. Это значит, что текущая строка является шапкой таблицы. То есть все лексемы в пределах ячейки, находящейся ниже дескриптора, являются его значением. Таким образом, значением дескриптора «Назва» является «Вода мин. Неаполис 0,33л». Аналогично заполняем значения «Од. вим.», «Кіл.», «Ціна без ПДВ», «Сума без ПДВ».

В табл. 3 однозначно отражены структура и содержимое документа, обработанного в соответствии с разработанным алгоритмом. Разработанный алгоритм позволяет автоматизировать процесс занесения информации из бумажных источников в электронные базы данных. Несмотря на то, что алгоритм требует дальнейшей детализации, возможность его применения на практике представлена на примере обработки счета-фактуры. Разработанная структура электронного документа обеспечивает полную идентичность электронной и бумажной версии документа.

Список литературы

1. Данилов А.Б. Офисный синдром / А.Б. Данилов, Ю.М. Курганова // Русский медицинский журнал. – 2011. – № 30. – С. 1902.
2. Роботизовані комп'ютерно-апаратні комплекси широкого призначення: необхідність і проблеми створення / Шевченко А.І. , Сальніков І.С., Сальніков Р.І., Цапко Є.В. // Искусственный интеллект. – 2012. – № 2. – С. 69-79.
3. Сальников И.С. Автоматизация деятельности офисного работника с помощью роботизированного компьютера / И.С. Сальников, С.В. Терещенко // Искусственный интеллект. Интеллектуальные системы ИИ-2013 : сборник материалов международной научно-технической конференции, Донецк, 2013.
4. Вся правда об OCR [Электронный ресурс]. – Режим доступа: http://blogerator.ru/page/ocr_abbyy-finereader-omnipage-readiris-tesseract.

References

1. Danilov A. B., Kurganova Y. M. Syndrome of office worker// Russian Medical Journal. – 2011. – № 30. – С. 1902.
2. Shevchenko A. I. Robotic computer-hardware complexes of wide application: the need and the problem of creating // Artificial intelligence. – 2012. - №2 – с. 69-79.
3. Salnikov I. S., Tereshchenko S. V. Automation of activity of office worker with a robotic computer. – Donetsk, Collection of the international scientific and technical conference «Artificial Intelligence. Intelligent Systems AI-2013», 2013.
4. The whole truth about OCR [Electronic resource]. – Access mode: http://blogerator.ru/page/ocr_abbyy-finereader-omnipage-readiris-tesseract.

RESUME

I.S. Salnikov, S.V. Tereshchenko

Algorithm of Entering Information from Paper Document s into Electronic Databases By Robotic Computer

Transferring information from paper sources in electronic databases of enterprises is one of the most important processes in document circulation. Relevance of automation of this process is caused by need to ensure the reliability of entered data, high speed document handling, as well as the fact that the volume of documents received for processing varies in different time periods.

While processing a document man catches sight a characteristic phrase and immediately understand what it refers, and how to handle this information. At the same time it is a very difficult task for a robot. One of the major challenges to the development of the robotic computer that can to replace an office employee in the workplace is pattern recognition.

Purpose of this work is to develop an algorithm of entering information from paper documents into electronic databases with a robotic computer and the formation of structure of an electronic document.

The developed algorithm allows to automate the process of entering information from paper documents into electronic databases. Despite the fact that the algorithm requires further of detail, the possibility of its application in practice is represented by the example of processing of the invoice. Developed structure of an electronic document provides a complete identity electronic and paper versions of the document.

Статья поступила в редакцию 14.04.2014.